Visual Analytics in Surveillance and Epidemiology: Challenges and Opportunities

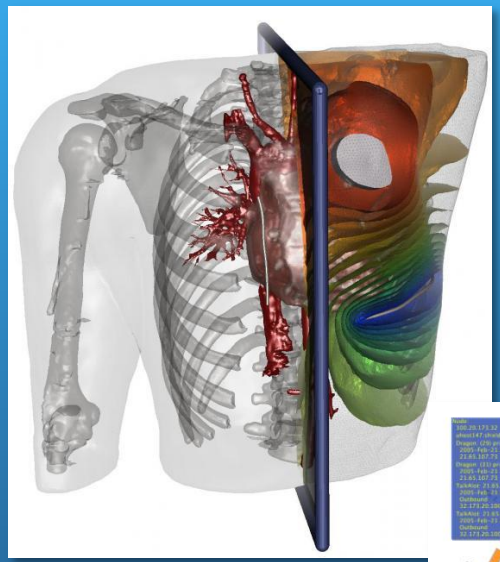
Yarden Livnat, PhD
Research Computer Scientist, SCI Institute, University of Utah
Matthew Samore, MD
Chief, Division of Epidemiology, VA Salt Lake, University of Utah

Rocky Mountain Center of Excellence in Public Health Informatics

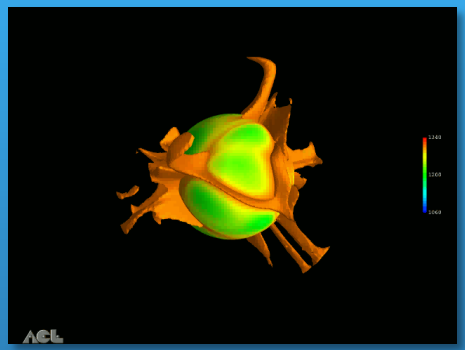
Outline

- Public health decision support
- Visual Analytics
- Challenges
  - Foodborne outbreak investigation domain
- Opportunities
  - *Epinome*, *CommonGround*
  - Graphs, Semantic

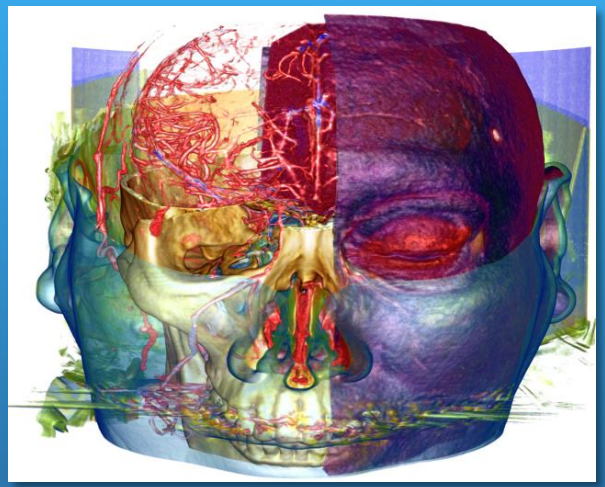
SCI Institute



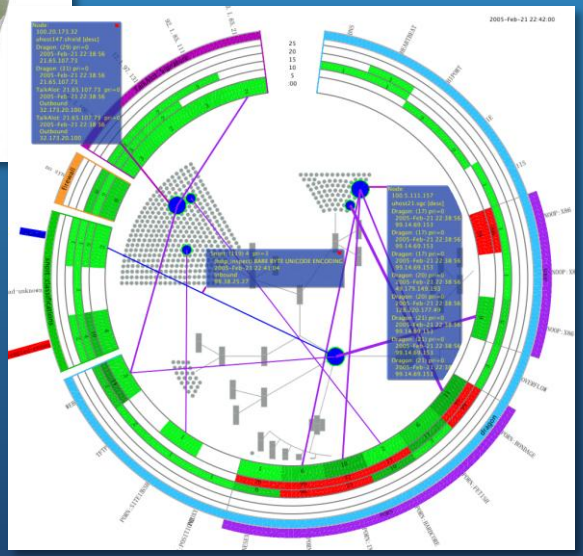
Biomedical Computation



Visualization of magma flow



Volume visualization



VisAlert: Network intrusion detection

Fire simulation and advanced rendering

Scientific Computing and Imaging Institute, University of Utah

www.sci.utah.edu

Public Health Decision Support

- About moving from patients to populations
- Our operational definition of public health decision support :
  - *"A system that integrates data with epidemiologic knowledge to assist decision-making by public health practitioners and officials to protect population health and prevent avoidable harm."*

What is a decision?

- A process which leads to a commitment to an action
- Components include: information, alternatives, attributes, values, preferences, goals
- “When choosing between two evils, always choose the one you haven't tried yet.”

  - Mae West

Our areas of translational research in public health decision support

- "Decision-making in the wild"
  - Describe heuristics and biases
    - Outbreak investigation
    - Use of protocols

- Dynamic information board
  - Probe cognitive processes that underlie diagnostic assessments and control decisions
  - How do epidemiologists use tools such as epidemic curves and maps

Areas of translational research, continued

- Simulation based decision support (current)
  - Contact tracing and pertussis
  - Community mitigation influenza
  - Active MRSA surveillance & decolonization
  - Control strategies for *Clostridium difficile infection*
  - Foodborne outbreaks (to be developed)

- Visual Analytics
  - Focus of today's presentation

Visual Analytics

- *The science of interactive visual interfaces to facilitate analytical reasoning*
- Goal:
  - Provide dynamic displays and interactions that support analytical reasoning
  - Facilitate the discourse between the user and the data

Visual Analytics: key concepts

- Focus on what the user wants to know rather than on what data is available
- Empower the user to explore the data from multiple aspects
- Facilitate parallel lines of thought
- Separation between the user and incidental form the data is stored

Visual Analytics: key concepts

- Don’t: data ➡ data transformation ➡ presentation
- Do: user ➡ question ➡ query ➡ data transformation ➡ presentation

Challenges

- We are not going to discuss all the challenges
  - That would be a challenge by itself
- There are many good and not so good approaches, solutions and systems
  - We are not going to discuss them either
- We will look at some of the challenges in one specific domain
- We will focus on those challenges that we have identified opportunities for from a user centric and visual analytics point of views

Foodborne Outbreak Investigation

1. The good news
   - Usually not many concurrent cases (in local context)
   - Can look at individual cases

   The bad news
   - Usually not many concurrent cases (in local context)
   - Hard to find correlations

Foodborne Outbreak Investigation

1. Usually not many concurrent cases
2. Many different variables
   - What kind of food they ate
   - Where was the food consumed
   - How was the food prepared
   - Who prepared the food
   - What ingredients may have been used
   - Where did the food/ingredients came from

Foodborne Outbreak Investigation

1. Usually not many concurrent cases
2. Many different variables
3. Many variables are outbreak specific
4. Incomplete data
   - Person did not know or remember
   - Question was not asked
   - Question may be asked later
5. Semi-structured data
6. Inconsistent data

Jumping through hoops

- Example of current workflow at a LHD
  - Start in TriSano (NEDSS)
  - Can't analyze the data in this system
  - Select and export data out of the system
    - Conversion to a single table
    - Lost of information
  - Cleaning
  - Create Excel spreadsheets
    - Summary of replies in a single cell
      - *E.g. had raw sushi at a sushi bar in Hawaii*
      - *E.g. Raw Fish Source: can't remember*
    - Free text form (lost of structured information)
    - No Standard vocabulary

Jumping through hoops

- Users have to fight their way through
- ETL (Extract, Transform, Load)
  - May not be available to the user
  - May not have been setup for the user tasks
  - User may not have the expertise to create and maintain
- ETL process is in the user's way it should be the consideration of only the software

Reasons for these hoops

- Data warehouses developed for internal use
- Relational databases are
  - Rigid
    - Everything has to fit into predefined structures
  - Require db expert knowledge
    - What each table/column actually means
    - Generate specific SQL queries
    - Create appropriate views for CUBE based systems
  - Close world assumption

Open vs. Close World Assumption

- Open world assumption

  The *truth-value* of a statement is independent of whether or not it is *known* to be true

- Closed world assumption

  Any statement that is not known to be true is false

Reasons for these hoops

- No good pre-defined queries or presentations
- No semantic
- What users need to know and correlate is very different for each investigation

Opportunities

1. Information Foraging
2. *Epinome*
3. Graph based vs. Relational databases
4. *CommonGround*
5. Semantic, ontologies and meta modeling

Opportunity 1

- Information Foraging Theory

Information Foraging Theory

- Analyze how humans collect information online
- Based on the analogy of wild animals gathering food
- An animal aims to optimize its effort
  - How long to stay in one patch before moving on?
    - How much food is available locally?
    - How hard is it to get to the next patch?
- Information foraging theory is the farmer point of view
  - How do I make bees come to my field and stay there?

Information Foraging Theory

- Cost-benefit analysis for navigation
  - How to get users to stay in your web pages
  - Opportunity: optimize public health personnel information foraging
- Information Scent
  - How much information (food) is available locally
  - Opportunity: reduce information overload
- Yarden's corollary

  If it's too hard to look for additional information users will keep analyzing the same old data

Opportunity 2

The discourse between the user and the data

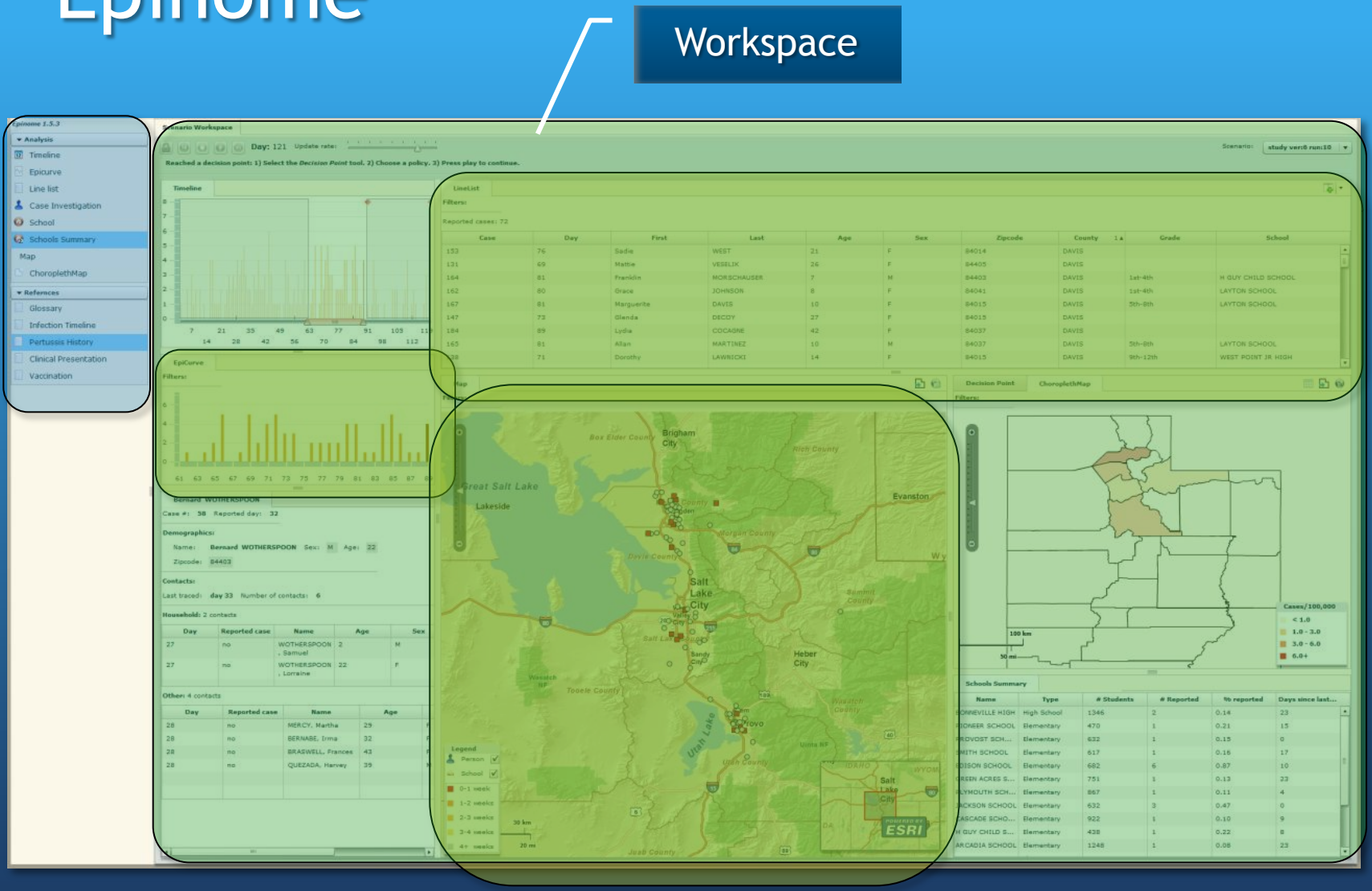
Or "*how to get the software out of the way*"

Epinome

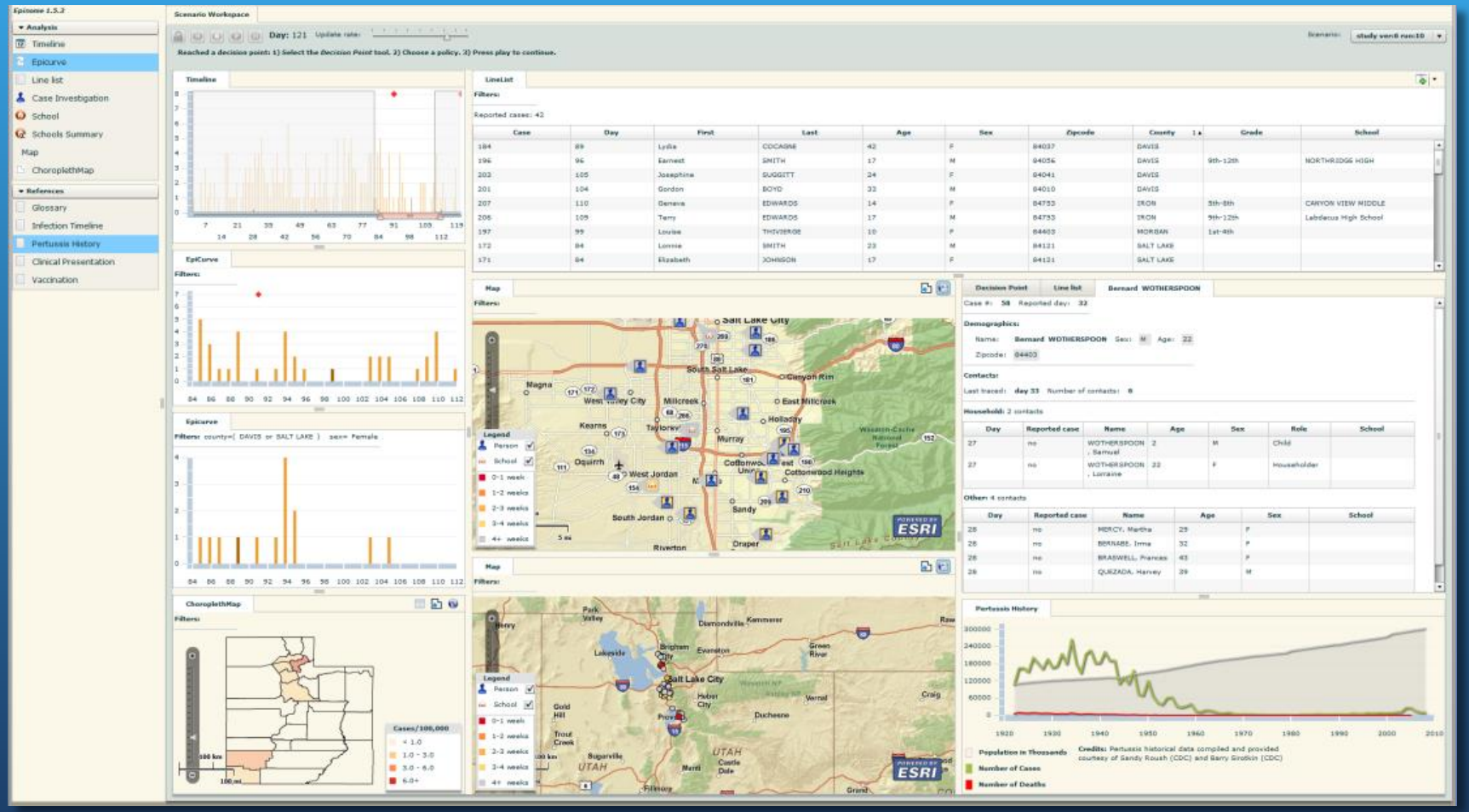
An interactive web-based visual analytic workbench

- Goal: to enable dynamic interactions that support analytical reasoning and facilitate the discourse between the user and the data
- Direct interaction with the data
- Support multiple lines of thought
  - More than drill down and level on demands

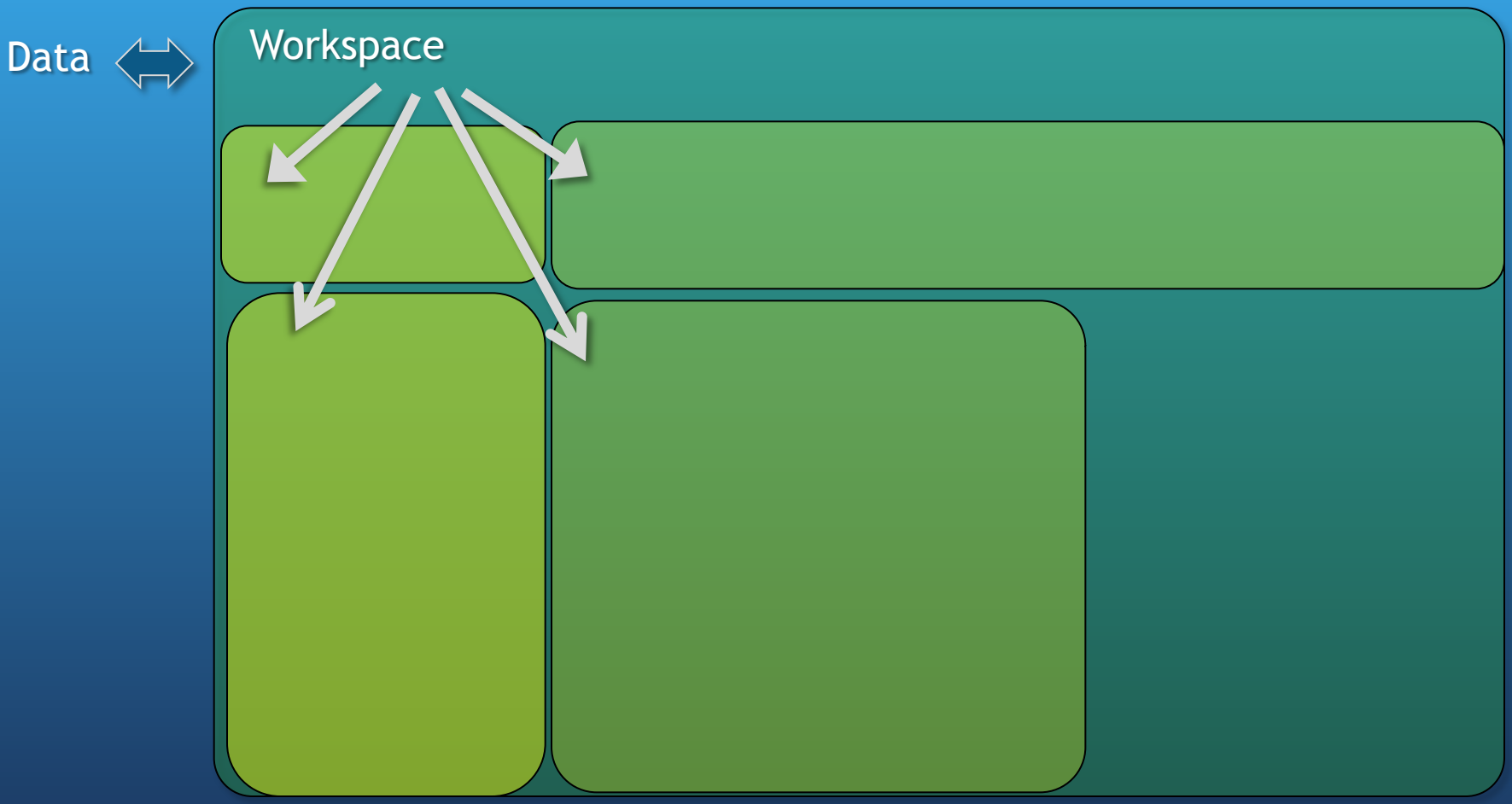
Epinome


Workspace
Epinome 1.5.3
Analysis
Timeline
Epicurve
Line list
Case Investigation
School
Schools Summary
Map
ChoroplethMap
References
Glossary
Infection Timeline
Pertussis History
Clinical Presentation
Vaccination
Scenario Workspace
Day: 121
Update rate:
Reached a decision point: 1) Select the Decision Point tool. 2) Choose a policy. 3) Press play to continue.
Scenario:
study ver:6 run:10
Timeline
LineList
Filters:
Reported cases: 72
Case
Day
First
Last
Age
Sex
Zipcode
County
Grade
School
EpiCurve
Filters:
Bernard WOTHERSPOON
Case #: 58 Reported day: 32
Demographics:
Name: Bernard WOTHERSPOON Sex: M Age: 22
Zipcode: 84403
Contacts:
Last traced: day 33 Number of contacts: 6
Household: 2 contacts
Other: 4 contacts
Map
Decision Point
ChoroplethMap
Filters:
Great Salt Lake
Lakeside
Brigham City
Box Elder County
Weber County
Davis County
Morgan County
Salt Lake City
Summit County
Tooele County
Heber City
Wasatch NP
Utah Lake
Provo
Utah County
Juab County
Uinta NF
Evanston
Wy
IDAHO
WYOM
Legend
Person
School
0-1 week
1-2 weeks
2-3 weeks
3-4 weeks
4+ weeks
30 km
20 mi
POWERED BY ESRI
Cases/100,000
< 1.0
1.0 - 3.0
3.0 - 6.0
6.0+
100 km
50 mi
Schools Summary
Name
Type
Students
Reported
% reported
Days since last...

Epinome a few minutes later…



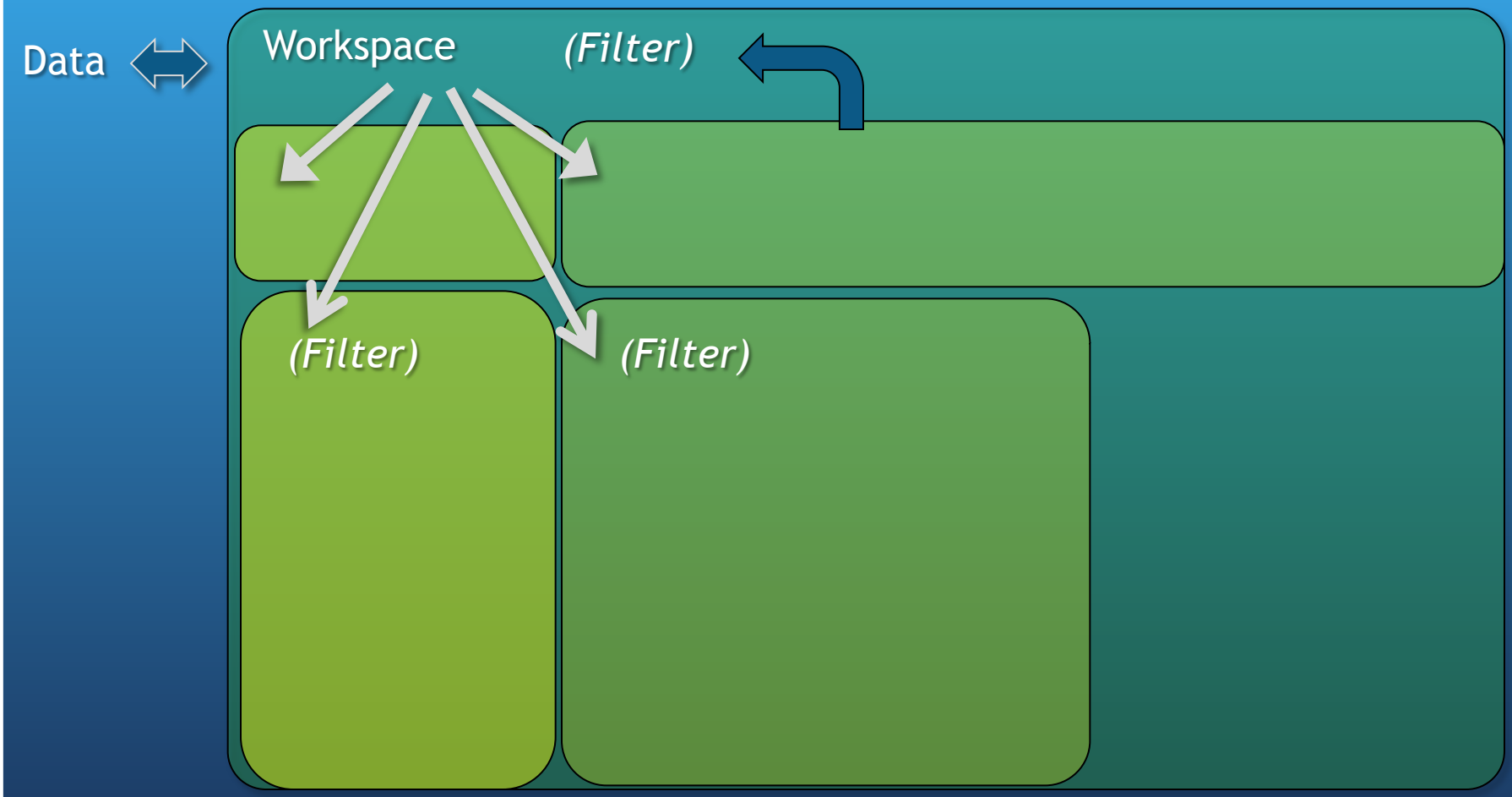
Workspace



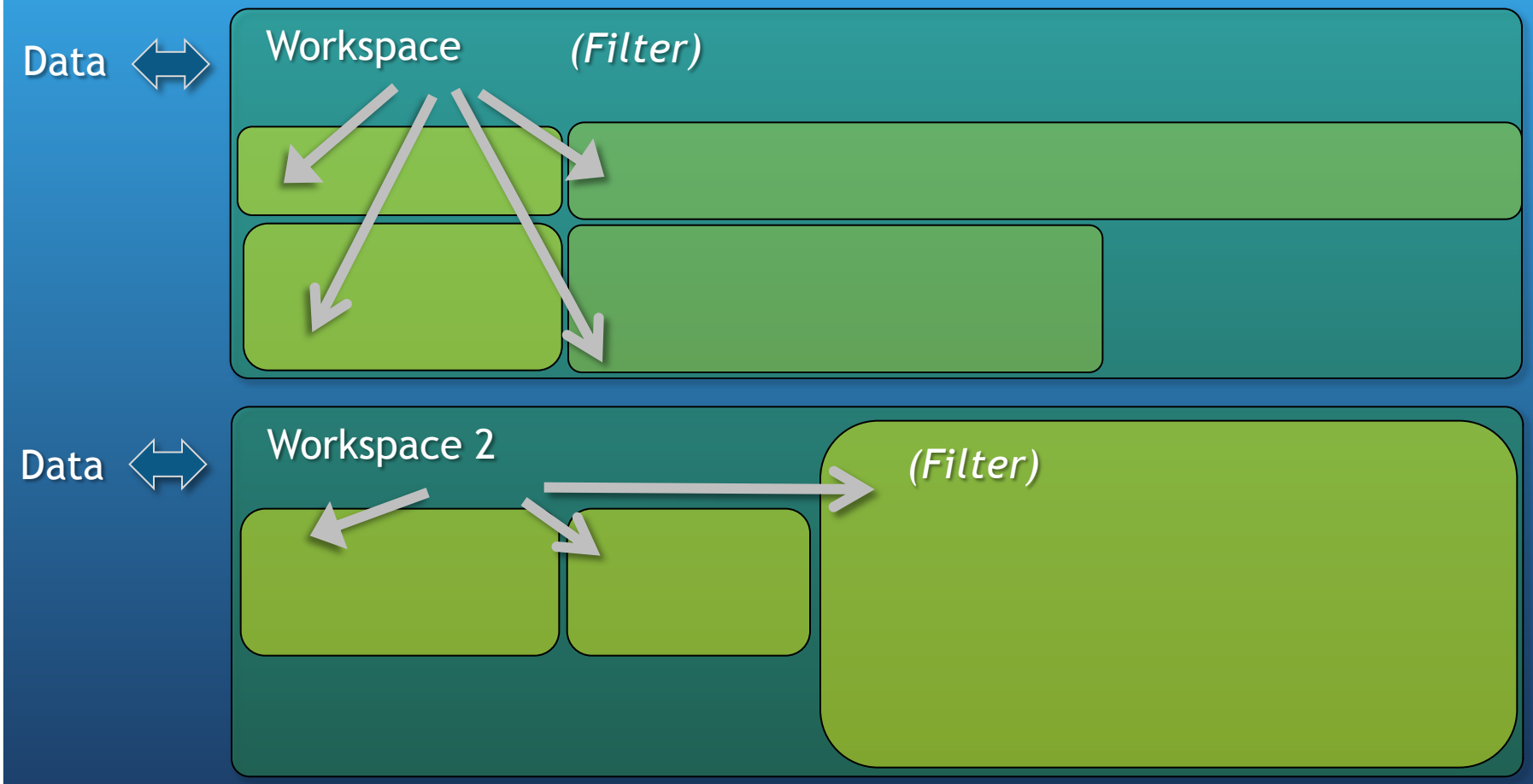
Epinome

- Demo

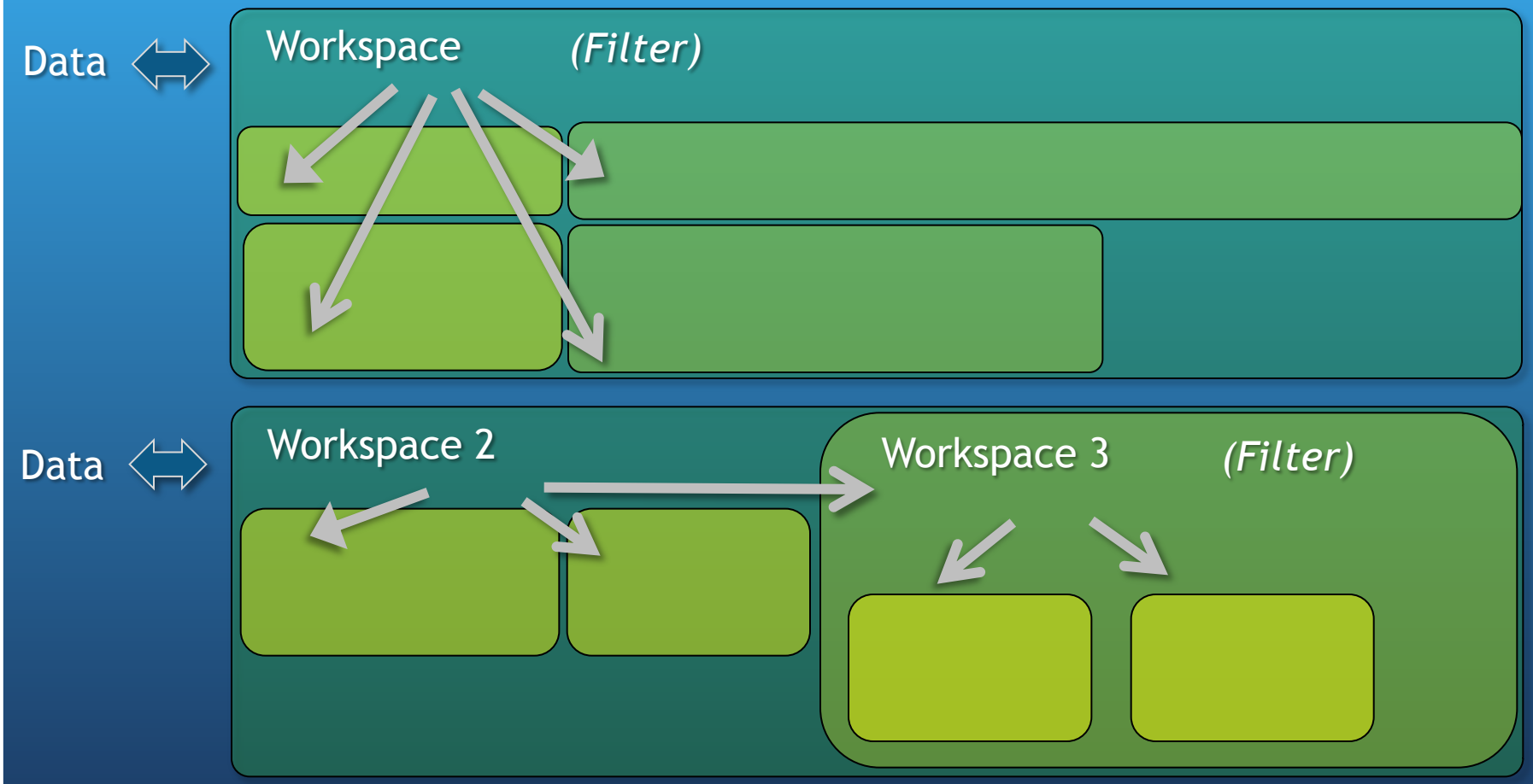
Workspace



Multiple Workspaces



Nested Workspaces



Epinome (summary)

An interactive web-based visual analytic workbench

- Multiple coordinated views
  - Loosely coordinated
  - Facilitate parallel lines of thought
- Direct user interactions
- Dynamically evolves to support the user workflow
- Multiple workspaces to support multiple hypothesis testing

Opportunity 3

How do users think about data?

Graph Based Approach

- Data is an unstructured collection of facts
- Facts are related to each others forming a graph
- John.address.zipcode.nearby.restaurants.serve.rawfish
- John…restaurants.serve(John.eat.food.raw)
- John.eat.food.raw.servedAt.restaurants.location.zipcode.nearbar.include(John.address.zipcode)

Graph Based Approach

- Graphs can adhere to the open world assumption
  - No need for “unknown” ethnicity
  - Adding information does not invalidate the graph
- Graphs better fit the human cognitive model
- Simple to augment
- Simple to drill down and details on demand

Graph Based Approach

- Breaking the relational model constraints
  - e.g. Person table, Address table
  - Artificial constructs to represent many-to-many relations
  - Close world assumption
    - e.g. Ethnicity (Hispanic, Not Hispanic, Other, Unknown)

Resource Definition Framework(RDF)

- RDF is a directed, labeled graph data format for representing information in the Web
- RDF triple
  - (subject, predicate, object)
- RDF Graph
  - A set of RDF triples
  - Each triple represent a single arc in a graph complete with beginning and ending nodes

SPARQL

A query language for querying RDF based graphs

- Can be used to express queries across diverse data sources
- Based around graph pattern matching.
- Contains capabilities for querying required and optional graph patterns along with their conjunctions and disjunctions.
- Read only, i.e. does not not modify the underlying RDF
- The results of SPARQL queries can be RDF graphs

SPARQL

Examples

1.
```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
SELECT ?name ?mbox
WHERE {
        ?x foaf:name ?name .
        ?x foat:mbox ?mbox .
}
```

2.
```
PREFIX dc10: <http://purl.org/dc/elements/1.0/>
PREFIX dc11: <http://purl.org/dc/elements/1.1/>
SELECT ?title
WHERE {
        { ?book dc10:title ?title }
        UNION
        { ?book dc11:title ?title }
}
```

Opportunity 3

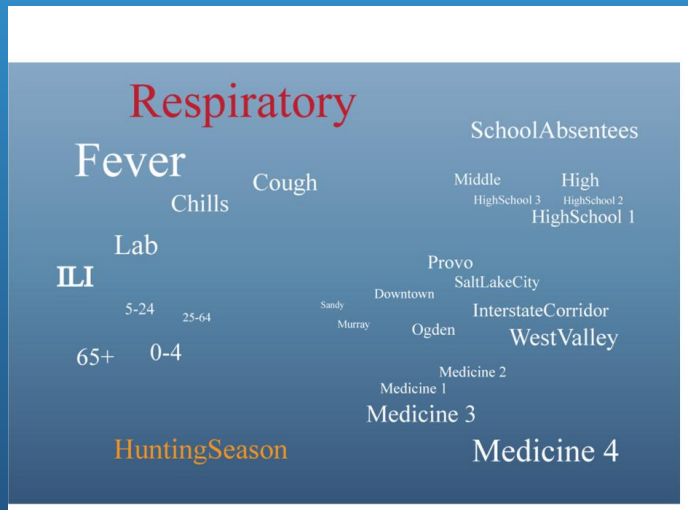
- Dynamic visual correlation using graph visualization

CommonGround

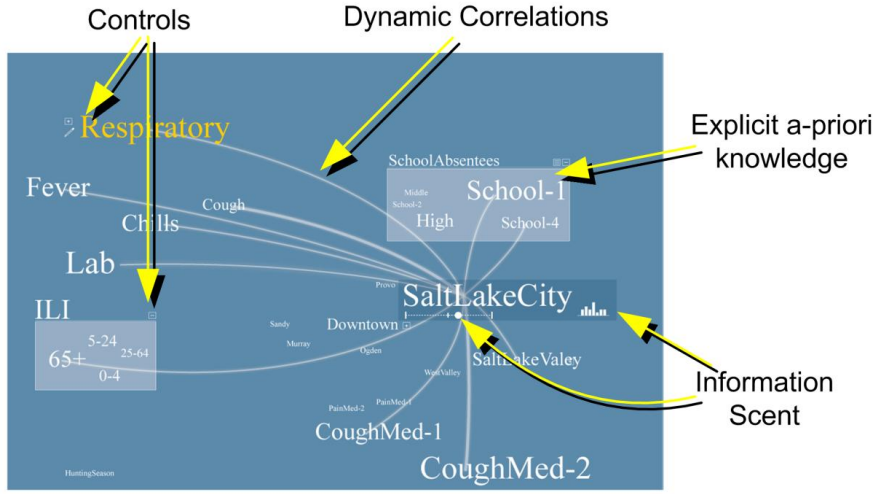
- Dynamic visual correlation
- Information overload
  - Graphs display do not scale
  - Too many nodes and edges (relationships)
- Information scent
  - Remove as much information as possible
  - Facilitate details on demand
  - Leave enough scent to help users find the most promising path

CommonGround

- Dynamic graph of current active concepts and current correlation between them



Situational Awareness



Visual correlation and Dynamic investigation

Opportunity 4

- Semantic

Semantics

- Controlled Vocabulary
  - A list of terms that have been enumerated explicitly
- Taxonomy
  + Organize terms in a hierarchical structure (e.g. *part-of*)
- Thesaurus
  + Add associated relationships (non-hierarchical *related-to*)
- Ontology
  + Formal logic-based language to specify the meaning of the terms
  + Constraints, e.g. cardinality, functional, transitive

  "An ontology is an explicit specification of conceptualization"

Semantics

- RDF: graph representation (syntax)
- RDFS: taxonomy
- OWL (Web Ontology Language)
  - A family of knowledge representation languages for authoring ontologies
  - Semantic reasoning
- Statements in an ontology can be grouped by
  - TBox: terminological statements, description of the domain
  - Abox: assertions
- Use of ontology
  - Users should work with assertions not the terminology
  - The visual analytics system should reference the TBox to extract semantic about the domain

Meta Modeling

- A meta model is an ontology that can be used to create a model of a domain
- Every meta model is an ontology but not every ontology is a meta model
- Meta modeling can facilitate development of software that is domain independent
- Semantic models for different public health domains
  - Foodborne outbreak investigation
  - IBIS – Indicator Based Information System

Acknowledgment

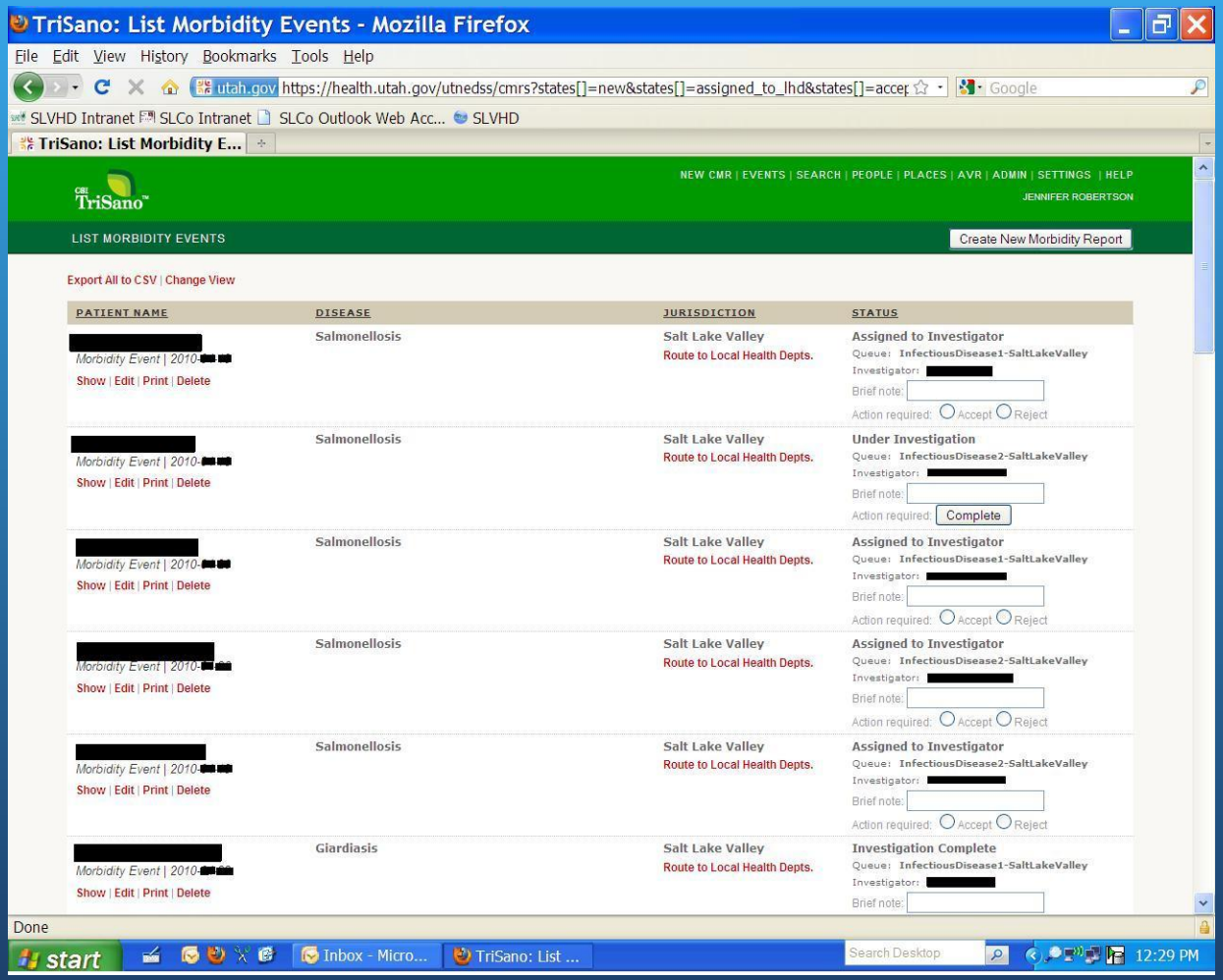
- Funding by CDC PO1 CD000284-01 & 5 PO1 HK000069-02
- U of Utah
  - Per Gesteland, Frank Drews, Warren Pettey
- Denver Health
  - Art Davidson
- Utah Department of Health
  - Robert Rolfs, Wu Xu, Rachel Herlihy, Jonathan Reid, Jonathan Anderson, William Lanier and Julia Hall
- Davis County HD
  - Brian Hatch, Cindy Burnett
- Salt Lake Valley HD
  - Jenny Robertson

Contact Information

- Matthew Samore Matthew.Samore@hsc.utah.edu
- Yarden Livnat yarden@sci.utah.edu

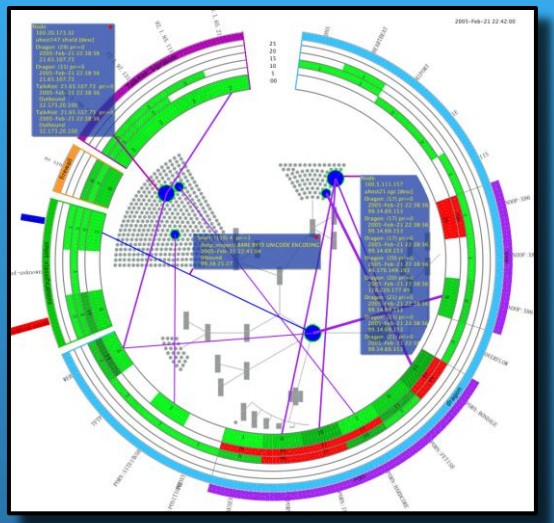
Extra Slides

Jumping through hoops

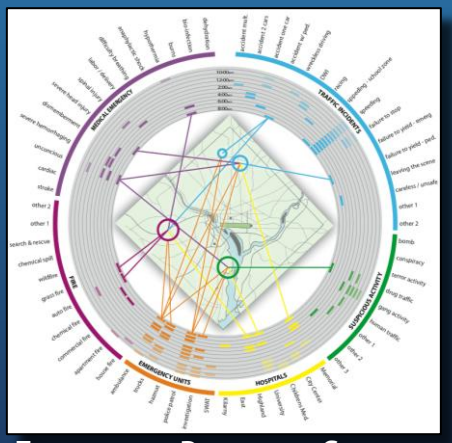


VisAware

- Visual Correlation for situational awareness
- Incorporates human judgment
- Global and Local perspectives
- Suggestive visualization
- Gradual disclosure of information
- Reduce information overload
- The *W premise*:
  - *What*
  - *When*
  - *Where*
  - *Who*



VisAlert: Network Intrusion Detection
ARDA and Air Force Research Lab



Emergency Response Center